Lab 2

February 15, 2022

General Instructions

1. You have to do this lab individually
2. You may use numpy, math, scipy, seaborn, matplotlib, pandas
3. All the code should be submitted in the form of a single Jupyter notebook itself.
4. Points for each sub-section are mentioned in the appropriate question.
5. The lab must be submitted on Google classroom. The code as well as the accompanying observations should be made part of the python notebook.
6. Code Readability is very important. Modularize your code by making use of classes, functions that can be flexibly reused wherever necessary. Also use self explanatory variable names and add comments to describe your approach wherever necessary.
7. You are expected to submit your inferences (preferably in a text block) and not just an error free code.
8. A readme file should precisely tell how to compile and run your program. Give the exact commands with respect to the datasets provided.
9. You should create a small video (hard limit max:15 mins), showing how you have solved the problems. A video will not be watched beyond the provided limit. The marks will be deduced accordingly.
10. You should start the video explaining the overall code like what functions it has and the overall logic. After explaining the video goes through the conceptual topics as mentioned specifically in each question.
11. You can use the software like free cam to record the video. Create different versions of the program for each variation (if your code organization is modular, this will be easier) and have the all the results ready before creating the video.
12. You should use the handbrake tool to reduce the overall size of the video. With hand-break the size of the video would be at max 50MB. That will help you and us to overcome the speed limitations.
13. Upload the video file separately with the ipynb file.
14. The marks will be given on the basis of quality of code, use of innovative data structures, scalability, correctness,inferences and your video explanations.
15. You are supposed to submit both the videos as well as code (ipynb file) on google classroom no later than **7th March 2022, 10 AM**. This is a strict deadline and any assignment submitted later will not be consider for evaluation.
16. Name the ipynb file as rollnumber-assignmentno.ipynb and the video file as rollnumber-assignmentno.mp4
17. Students are expected to follow the honor code of the class. We will follow a strict anti-plagiarism policy.

1 Logistic Regression / LDA

For this section we will be using a beginner friendly dataset to test a simple binary classification task of whether a person will be interested in learning a new sport based on just two attributes namely, age and interest quotient, that can be downloaded from Kaggle. With respect to this dataset, implement/report the following:

1. Plot the dataset using different colors for the two classes. [5 Marks]
2. Implement the least square method for classification and plot the decision boundary. Clearly describe your results. Is the decision boundary able to classify the points correctly? [15 Marks]
3. Implement the logistic regression using gradient descent method. Choose the initial values of w in the range $[-0.1, 0.1]$. Plot a 3D figure depicting the sigmoid function obtained along with the same color coding of the points. Did the performance improved as compared to previous question? [15 Marks]
4. Plot the decision boundary obtained for logistic regression. [5 Marks]
5. Find the linear discriminant boundary and describe your results. [10 Marks]
6. Logistic regression considers only linear decision boundaries. One way to go from linear decision boundaries to non-linear decision boundaries is by considering polynomial curve of higher degree. For example, if input attributes are x_1, x_2 then transforming it into 2 degree polynomial will give features: $\{x_1, x_2, x_1^2, x_2^2, x_1x_2, 1\}$. Identify an appropriate degree of the transformation that results in the optimal performance via logistic regression. Clearly explain your choice. [10 Marks]
7. Above expansion will result in non-linear decision boundary. Plot the boundary along with the dataset points. [5 Marks]

2 PCA / Decision Trees/ Random Forests

The dataset we used in the previous section had just two numeric attributes. In this section we will look at a slightly sophisticated dataset having a mix of numeric and categorical attributes describing an adult. The dataset can be downloaded from UCI Machine Learning repository. The task is to predict whether the person defined by the given set of attributes earns more than 50000 or less (Binary classification task). Implement the following and state your results with respect to this dataset.

1. Implement the decision tree algorithm to classify whether the income of a particular user exceeds \$50K per year or not. Divide the data into two sets: Training set (80%) and validation set (20%). Plot the training error and validation error against the number of nodes present in the decision tree. Describe the optimal decision tree in your video. [15 Marks]
2. Create 10 datasets using bootstrap technique and rerun the part 1 to find the optimal decision tree for each of these datasets. Report the final error by taking the average of each decision tree and report your findings. Did the performance improved? [10 Marks]
3. Implement PCA to find optimal number of features. Plot the error of optimal decision tree against the number of features. How many features did it require to match the performance of the tree obtained in the first part. [10 Marks]